Journal of Applied and Theoretical Social Sciences

JATSS, 2026; 8(1), 189-212

First Submission:04.01.2026
Revised Submission After Review:05.03.2026
Accepted For Publication:24.03.2026
Available Online Since:25.03.2026

Research Article

The Power of Words for Credit Risk Classification: Sentiment Analysis As An Alternative Credit Rating Tool

Yunus Emre Akdoğan[a]

Abstract

Introduction: This study examines whether qualitative narratives in credit rating reports can be systematically structured using machine learning and to what extent indicators derived from these narratives carry information for predicting national credit ratings. The research is significant because it operationalizes narrative-based assessments, central to rating processes yet difficult to integrate into quantitative models, making them measurable and analytically tractable.

Method: This study analyzes publicly available institutional credit rating reports. Reports statements are first classified as strengths or constraints using a transformer-based model and then transformed into narrative indicators. Using only these features, national credit ratings are predicted via multi-class classification, with firm-level train–test splitting to avoid data leakage. Performance is assessed using accuracy, precision, recall, and F1-score.

Results or Findings: The strength–constraint classifier achieved high performance (accuracy = 0.977; F1 = 0.981). For rating prediction, accuracy was 0.517, with macro-F1 = 0.403 and weighted-F1 = 0.500. Class-level results were more stable for AA and BBB, but limited for AAA and BB.

Discussion or Conclusion: The findings indicate that indicators derived from report narratives carry a meaningful signal about credit ratings, yet this signal is not sufficient on its own. The proposed approach can be used for rapid summarization of newly released reports, tracking changes in the Strengths–Constraints balance, and supporting scalable credit surveillance and early-warning systems. Future research is encouraged to adopt approaches that integrate narrative indicators with financial and macroeconomic variables.

Keywords: credit rating reports, credit rating prediction, transformer-based classification, financial natural language processing

JEL Codes: G24, G17, C45, C55

[a] Assist. Prof. Dr., Yozgat Bozok University, Faculty of Economics and Administrative Sciences, Department of Business Administration, Yozgat/Türkiye, emre.akdogan@bozok.edu.tr, ORCID ID: https://orcid.org/0000-0002-1761-2869, (Corresponding Author)

Journal of Applied and Theoretical Social Sciences

JATSS, 2026; 8(1), 189-212

İlk Başvuru: 04.01.2026
Düzeltilmiş Makalenin Alınışı: 05.03.2026
Yayın İçin Kabul Tarihi: 24.03.2026
Online Yayın Tarihi: 25.03.2026

Araştırma Makalesi

Kredi Riski Sınıflandırmasında Kelimelerin Gücü: Alternatif Kredi Derecelendirme Aracı Olarak Sentiment Analizi

Yunus Emre Akdoğan[a]

Öz

Giriş: Bu çalışma, kredi derecelendirme raporlarında yer alan nitel anlatıların makine öğrenmesi ile sistematik biçimde yapılandırılıp yapılandırılamayacağını ve bu anlatıdan türetilen göstergelerin ulusal kredi notunu öngörmede ne ölçüde bilgi taşıdığını incelemektedir. Araştırma, derecelendirme süreçlerinde merkezi rol oynayan ancak çoğu zaman nicel modellere doğrudan entegre edilemeyen anlatı temelli değerlendirmeleri ölçülebilir ve analitik olarak işlenebilir hale getirmesi bakımından önemlidir.

Yöntem: Bu çalışma, kamuya açık kurumsal kredi derecelendirme raporlarını incelemektedir. Raporlardaki ifadeler transformer tabanlı bir modelle güçlü yönler ve kısıtlar olarak sınıflandırılarak anlatı göstergelerine dönüştürülmüştür. Yalnızca bu özellikler kullanılarak ulusal kredi notu çok sınıflı bir problem olarak tahmin edilmiş, veri sızıntısını önlemek için firma bazlı ayrım uygulanmış ve performans accuracy, precision, recall ve F1-skoru ile değerlendirilmiştir.

Sonuçlar ya da Bulgular: Güçlü yönler-kısıtlar sınıflandırması yüksek performans üretmiştir (accuracy=0,977; F1=0,981). Ulusal kredi notu tahmininde ise örneklem dışı doğruluk 0,517; macro-F1 0,403; weighted-F1 0,500 olarak elde edilmiştir. Sınıf bazında AA (F1=0,605) ve BBB (F1=0,591) sınıflarında görece daha istikrarlı sonuçlar görülürken, AAA'da düşük duyarlılık ve BB'de düşük destek nedeniyle zayıf performans dikkat çekmektedir.

Tartışma ya da Yapılan Çıkarımlar: Bulgular, rapor anlatılarından türetilen göstergelerin kredi notuna ilişkin anlamlı ancak tek başına yeterli olmayan bir sinyal taşıdığını göstermektedir. Önerilen yaklaşım, yeni yayımlanan raporların hızlı özetlenmesi, güçlü yönler-kısıtlar dengesindeki değişimlerin izlenmesi ve ölçeklenebilir kredi gözetimi/erken uyarı sistemleri için kullanılabilir. Gelecek araştırmaların, anlatı göstergelerini finansal ve makro değişkenlerle birleştirmesi gibi yöntemler önerilmektedir.

Anahtar Kelimeler: kredi derecelendirme raporları, kredi notu tahmini, transformer tabanlı sınıflandırma, finansal doğal dil işleme

JEL Kodlar: G24, G17, C45, C5

[a] Dr. Öğr. Üyesi, Yozgat Bozok Üniversitesi, İktisadi ve İdari Bilimler Fakültesi, İşletme Bölümü, Yozgat/Türkiye, emre.akdogan@bozok.edu.tr, ORCID ID: https://orcid.org/0000-0002-1761-2869, (Sorumlu Yazar)

Introduction

Accurate and timely measurement of credit risk is a fundamental requirement for financial institutions, investors, and regulators. In recent years, the credit risk literature has developed a broad set of "credit scoring/classification" techniques, ranging from traditional statistical approaches to machine learning methods (Baesens et al., 2003; De Oliveira & Basso, 2025; Pai et al., 2015; Siham et al., 2021; Teles et al., 2021; Thomas, 2000; Yobas et al., 2000). As the scale of credit markets has expanded, the fact that even small improvements in predictive accuracy can generate substantial economic gains through pricing, provisioning, capital allocation, and risk management has increased interest in benchmarking alternative algorithms. Indeed, numerous comparative studies show that modern machine learning approaches often outperform conventional baseline models across various credit scoring settings (Darwish, 2025; Lessmann et al., 2015; Machado et al., 2025; Machado & Karray, 2022; Marqués et al., 2012; Paleologo et al., 2010; Tripathi et al., 2019; Tsai, 2014; Xu et al., 2025).

However, a key element of real-world credit assessments, namely the qualitative narratives that accompany credit decisions, remains relatively underutilized in empirical modeling. In practice, credit risk decisions are communicated not only through a score or rating but also through textual explanations that summarize the rationale behind the assessment. Credit rating reports, in particular, often articulate the basis of the evaluation under headings such as *strengths* and *constraints*. Despite their central role in how stakeholders interpret credit risk, predictive models still predominantly rely on structured financial ratios and other tabular variables.

In corporate credit rating settings, this gap is particularly pronounced. Rating reports routinely articulate the main factors underlying the assigned grade as rating drivers, typically organized under sections resembling *strengths* and *constraints*. These narratives are a core part of how the rating decision is justified and communicated to stakeholders. However, turning such text into scalable, quantitative inputs is difficult because the content is unstructured, varies across reports, and relies on domain-specific phrasing. Recent advances in natural language processing, especially transformer-based architectures, make it increasingly feasible to parse these narratives systematically and derive measurable indicators from them. Even so, fully integrated studies that convert rating narratives into structured variables and evaluate their explanatory and predictive value for rating outcomes under leakage-robust designs remain relatively scarce.

This study focuses on the question: "Can qualitative descriptions in credit rating reports be systematically transformed into structured indicators through machine learning, and can these indicators provide statistically significant information in predicting national credit ratings?" In this context, the proposed two-stage framework aims to contribute to addressing this gap by systematically analyzing these narratives. In the first stage (Task 1), statements in rating reports are classified into *strengths* and *constraints* using a fine-tuned transformer-based classifier. In the second stage (Task 2), these outputs are aggregated at the report level to construct narrative-based indicators, such as the counts and balance of *strengths* and *constraints* as well as composite indices, and national credit rating classes are then predicted using a multi-class model based solely on these text-derived features. To reflect realistic deployment conditions and prevent information leakage, the train–test split is performed at the firm level, ensuring that reports from the same company do not appear in both the training and test sets.

The empirical findings draw a sharp distinction between a task that is linguistically learnable and a prediction task necessitating a more comprehensive information set.

Specifically, while the classification of *strengths* and *constraints* achieves near-perfect performance (Accuracy≈0.98; F1≈0.98), utilizing narrative-derived indicators alone for rating class prediction results in only moderate success (Accuracy≈0.52; macro−F1≈0.40), with significant performance drops observed in sparse or extreme classes. These results suggest that while rating narratives provide meaningful insights, they offer an incomplete signal regarding the final credit rating. Consequently, although text-based models excel at the scalable extraction and summarization of key credit drivers, they cannot fully replace quantitative fundamentals and the broader data environment used in the formal rating process.

This study contributes to the literature in three main ways. Methodologically, it employs machine learning to systematically transform qualitative credit rating report narratives into measurable indicators, thereby bridging the gap between narrative-based assessments and quantitative modeling, and establishing a reproducible and auditable feature space by modeling rating narratives at the granular level as *strengths* and *constraints* and synthesizing them into report-level indicators. From a data-driven perspective, the study integrates unstructured text data directly from rating reports into empirical analysis and predictive models, using a firm-level, leakage-robust evaluation framework to delineate the predictive boundaries of text-derived signals and identify conditions under which narrative-based modeling is insufficient. Empirically, by applying the framework to a national credit rating sample, the study tests the predictive power of sentiment-oriented indicators such as *strengths* and *constraints*, providing novel and concrete evidence, and offering a practical foundation for scalable credit monitoring through rapid processing of new reports and quantification of shifts in credit quality, while highlighting the importance of integrating complementary quantitative variables for robust early-warning analytics.

In the remainder of the paper, Section 2 reviews the literature on credit scoring and risk analytics. Section 3 describes the dataset, the models, the performance metrics used in the study and discusses the empirical findings. The final section concludes the paper by highlighting implications for credit monitoring and outlining directions for future research.

Literature Review

Credit risk assessment and rating/score prediction have long been central topics in finance and risk management, traditionally addressed with statistical classification models based on firm- and borrower-level quantitative indicators. Over the past two decades, rapid growth in computational power and data availability has shifted this literature toward machine learning approaches that can capture nonlinear relationships, complex interactions, and heterogeneous patterns in default and creditworthiness signals. Accordingly, a substantial body of research has benchmarked a wide range of algorithms, including logistic regression, decision trees, support vector machines, and neural networks, while a parallel stream has emphasized performance gains through ensemble learning, feature engineering, and robust evaluation under class imbalance and asymmetric error costs. More recently, the field has begun to extend beyond purely numerical predictors by incorporating unstructured text, such as disclosure narratives and analyst or rating-agency communications, to quantify qualitative information embedded in credit-related documents. Building on this broader trajectory, the following section reviews key methodological advances in credit scoring and related predictive frameworks, with particular attention to ensemble strategies and the emerging role of text-based signals.

In this context, ensemble learning methods have attracted particular attention for their potential to improve generalization performance and reduce classification errors. In their study,

Wang et al. (2011) compared the effectiveness of Bagging, Boosting, and Stacking ensemble methods using four base learners (Logistic Regression Analysis, Decision Tree, Artificial Neural Network, and Support Vector Machine) for credit scoring. The study revealed that the ensemble methods notably improved the performance of individual base learners. Specifically, Bagging outperformed Boosting across all credit datasets, and both Stacking and Bagging showed superior performance in terms of Decision Tree, average accuracy, Type I error, and Type II error.

Liu & Schumann (2005) conducted an empirical study to investigate how various algorithms, including "ReliefF", "Correlation-based", "Consistency-based", and "Wrapper", can enhance the performance of credit scoring models in terms of model simplicity, speed, and accuracy. They used real data sets and applied four classification algorithms: "model tree (M5)", "back-propagation multilayer perceptron", "logistic regression", and "k-nearest neighbor". The study findings revealed that the consistency-based and wrapper feature selection methods outperformed the other two methods. The learning curves of the consistency-based and wrapper methods displayed lower error rates in the early range (from 0 to approximately 33 features), indicating that their models achieved lower error rates earlier than the other two methods. Upon comparing the algorithms, it was noted that the k-nearest neighbor algorithm exhibited improvements in model accuracy, while no improvement was observed for the other algorithms. Consequently, the researchers concluded that reducing the number of features enhances classification accuracy, reduces training time, and simplifies the final models.

Teles et al. (2020) conducted a study to compare the effectiveness of fuzzy sets and neural network-based decision trees in credit scoring to estimate the recovered value using a dataset of 1890 borrowers. The study concluded that both models allow uncertainty modeling in credit scoring. Specifically, the study found that fuzzy logic is more adept at modeling uncertainty, while the decision tree model is better suited for addressing the problem at hand.

Trivedi (2020) utilized German credit data in their study to evaluate credit risk and focused on developing a credit scoring prediction model using artificial intelligence. The research included a comparative analysis of evaluation metrics to identify the most effective combination of machine learning classifier and feature selection technique. The study compared Bayesian, Naïve Bayes, SVM, C5.0, and Random Forest methods as machine learning classifiers, as well as Chi-Square, Information-gain, and Gain-Ratio as feature selection algorithms. Ultimately, the analysis concluded that the combination of Random Forest and Chi-Square is the most optimal pair for creating credit scoring models.

In their study, Dumitrescu et al. (2022) developed a penalized logistic tree regression (PLTR) model, which combines decision trees with logistic regression. The model was designed to address the lack of interpretability in many existing machine-learning algorithms used for credit scoring. Through Monte Carlo simulations and the application of the PLTR model to four real credit scoring datasets, they demonstrated its interpretability and predictive power. Their findings show that the PLTR effectively captures non-linear effects in credit scoring data while remaining interpretable. Empirical evidence indicates that the PLTR significantly improves credit risk prediction compared to logistic regression and competes closely with the random forest method. Furthermore, they assessed misclassification costs using expected maximum profit analysis and concluded that the PLTR method considerably reduces misclassification costs.

In a recent study, Bucker et al. (2022) sought to address the limitations of traditional credit scoring methods, which often result in higher reserves or increased loan defaults. Their

goal was to introduce a transparent and auditable machine learning model that not only delivers accurate risk estimation but also ensures understandability. The study emphasized the importance of transparency, auditability, and explainability in machine learning models for credit scoring. Interestingly, the researchers found that the baseline scorecard performed remarkably well compared to advanced machine learning techniques such as Gradient Boosting or Support Vector Machines. They also predicted that with the anticipated increase in the use of transactional and external data sources, credit scoring data will become more complex, underscoring the growing significance of attribute engineering.

In a study conducted by Gambacorta et al. (2024), the effectiveness of credit-scoring models based on machine learning techniques was compared to that of traditional loss and default models. The study utilized proprietary transaction-level data from a leading fintech company in China, spanning the period between May and September 2017. The performance of different models in predicting losses and defaults during both normal economic conditions and economic shocks was evaluated. Specifically, the study analyzed the impact of a regulatory policy change on shadow banking in China, which resulted in reduced lending and deteriorating credit conditions. The findings revealed that the machine learning-based model, incorporating non-traditional data, demonstrated superior predictive performance for losses and defaults relative to traditional models in response to a negative shock to aggregate loan supply.

In their study, Tripathi et al. (2020) underscored the importance of accurately distinguishing between "creditworthy" and "non-creditworthy" categories to enhance the performance of credit scoring models, especially for the unreliable group, thus affecting the profitability of financial institutions. Their investigation aimed to assess the influence of various combinations of feature selection and classification approaches. They applied nine feature selection methods and various classification approaches to datasets such as Australian, Bank-marketing, Bankruptcy, Japanese, German-categorical, German-numerical, and Taiwanese. Their findings indicated that TDNN (Time Delay Neural Network) and RF showcased the best and second-best performance across most of the credit scoring datasets. Furthermore, the study revealed that the Correlation Coefficients-based Feature Ranking (CFS) approach consistently delivered superior results across most of the classification algorithms.

In their study aiming to develop a model for evaluating personal loans based on big data from the lending club dataset, Wu & Pan (2021) employed the pdC-RF algorithm to optimize data feature correlation and reduce the dimensionality of personal loan data from 145 to 22 dimensions. They evaluated the dataset using random forest, support vector machine, and logistic regression models, and utilized weight of evidence (WOE) coding to measure the probability of a grouping in the features being predicted as a negative example. Following the model performance comparison, they concluded that logistic regression is more suitable for the personal loan assessment model.

In their study, Gunnarsson et al. (2021) sought to examine the effectiveness of deep learning algorithms for credit scoring. They developed and compared two deep learning architectures, namely a multilayer perceptron network and a deep belief network, with two traditional and two ensemble methods for credit scoring. The performance of these models was then evaluated using various credit scoring datasets and performance metrics. Furthermore, the researchers compared Bayesian statistical testing procedures with frequentist non-parametric testing procedures, which are conventionally considered best practices in credit scoring. The traditional methods considered Logistic Regression and Decision Tree, while the ensemble learning methods comprised Random Forest and XGBoost. For deep learning, the models used were Multilayer Perceptron Networks and Deep Belief Networks. The analysis of these

different classification algorithms for credit scoring led to two principal findings. Firstly, the XGBoost algorithm, an ensemble method, exhibited the best performance for credit scoring among all the methods considered. Secondly, it was found that deep neural networks did not outperform their shallower counterparts and were considerably more computationally intensive to construct. Consequently, based on this comparison, the researchers suggested that deep learning algorithms may not be the most suitable models for credit scoring and recommended the XGBoost method as the preferred approach for credit scoring activities due to its superior classification performance.

Boughaci & Alkhawaldeh (2020) undertook a study to compare various machine learning techniques using datasets from financial institutions in six different countries and the "Give Me Some Credit" dataset. The objective was to identify the most effective methods for credit scoring and insolvency prediction. Their findings showcased machine learning methods capable of distinguishing between suitable and unsuitable applicants or companies by generating scores for them. However, they determined that no single method consistently outperformed the others across all datasets. Furthermore, they noted significant variations in the performance of the methods across different datasets. For instance, the Bayesian net method proved most effective for the German and Give Me Some Credit datasets, while the LogitBoost method outperformed others for the Polish and Australian datasets, the AdaBoost method for the Japanese dataset, and the Random Forest method for the Taiwan dataset. In the Indian Qualitative Bankruptcy dataset, nearly all methods demonstrated comparable performance due to the unique nature of the data.

In their study, Radovanovic & Haas (2023) aimed to enhance traditional bankruptcy prediction models by integrating the socio-economic consequences of their forecasts alongside the prediction of the bankruptcy event itself. They leveraged a substantial real-world dataset comprising over 190,000 company-year observations of listed North American companies between 1985 and 2020. The study compared the performance of various machine learning algorithms, including logistic regression, support vector machines, random forest, and neural networks, with traditional statistical models like linear discriminant analysis. The results consistently demonstrated the superior performance of the machine learning models over the traditional statistical models. The study emphasized the importance of considering the financial and socio-economic implications when selecting the most appropriate bankruptcy forecasting model, as even minor disparities in model performance can have significant ramifications in these domains.

A study conducted by Toudas et al. (2024) focused on analyzing bankruptcy prediction within the construction industry in Greece during an economic crisis. The research utilized financial data from construction companies listed on the Greek Stock Exchange in Athens to compare the performance of three distinct models for predicting corporate insolvency, namely various iterations of the Altman Model, the Ohlson Model, and the Zmijewski Model. The objective was to assess the predictability of these models and to ascertain their efficacy in predicting bankruptcy for insolvent companies. The results indicated that both the original Altman model and its revised versions exhibited low overall predictability for the three years leading up to bankruptcy.

In 2023, Máté et al. conducted a study to determine the most effective method for predicting business failure in non-financial firms in Pakistan. They employed various machine learning models on 36 financial ratios to address questions regarding model selection and the best ratios to use. The study highlighted several financial ratios, such as return on assets, operating return on assets, debt coverage ratio, asset turnover, earnings per share, debt/asset

ratio, cash return on assets, and quick ratio, as valuable for predicting bankruptcy. The findings showed that decision tree, AdaBoost, and gradient boosting models performed exceptionally well with 100% accuracy, while SVM and logistic regression models demonstrated flexibility in feature selection, achieving accuracy rates between 89% and 99%. Conversely, the Naive Bayes model performed inadequately, yielding an accuracy range of 58% to 70%.

In 2023, Gajdosikova & Valaskova endeavored to develop a model for predicting bankruptcy based on the financial data of 3,783 Slovak enterprises in the manufacturing and construction sectors in 2020 and 2021. Utilizing multiple discriminant analyses, they identified key financial indicators for the prediction. The results emphasized the self-financing ratio as the most accurate variable in the model. Ultimately, the model, developed through multiple discriminant analyses, exhibited an impressive overall discriminant accuracy of 93%.

In 2022, a study by Muñoz-Izquierdo et al. explored the impact of specific sections within extended audit reports on corporate credit ratings. The researchers employed four machine learning techniques - C4.5 decision tree, PART algorithm rule classifier, rough set methodology, and logistic regression. Their findings demonstrated that accurately identifying key audit matters within the report enabled the evaluation of credit scores with 74% accuracy, using the rules provided by the PART algorithm. Additionally, internal and external key audit matters influenced a company's credit rating disclosure. The study also highlighted the similar predictive power of rule induction classifiers. Interestingly, when combining audit data with accounting ratios, the predictive accuracy of the model increased to 84%, surpassing that of the existing literature.

Issa et al. (2024) examined the probability of bankruptcy across 20 financial sector institutions by analyzing indicators such as liquidity, profitability, debt composition, and operational efficiency derived from financial statements spanning 2000 to 2018. These metrics were juxtaposed with regulatory standards and evaluated for their low, medium, or high-risk implications, culminating in an overarching risk assessment. Additionally, the model incorporates an algorithm that bolsters the dependability of the risk evaluation by accounting for excessive debt levels. The findings underscore that excessive debt adversely affects profitability, leading to lower stock returns and a greater likelihood of bankruptcy. Furthermore, excessive debt and leverage at major financial institutions pose a risk of systemic risk, potentially triggering a domino effect across the global financial system. These insights carry practical implications for investors and stakeholders, furnishing enlightening perspectives to aid decision-making processes, especially during periods of economic volatility.

Slapnik and Lončarski (2023) examine the determinants of sovereign credit ratings by applying textual analysis to credit rating reports. The study uses sentiment and subjectivity scores extracted from S&P, Fitch, and Moody’s reports, representing, respectively, the general perception of a country and the qualitative judgment of the rating committee. The findings indicate that soft information (objectively unobservable factors affecting a country’s debt repayment capacity) and bias proxies provide significant additional insight in predicting sovereign credit ratings. Differences were observed between emerging and advanced markets, with emerging markets exhibiting a higher reliance on the rating committee’s qualitative judgment. Moreover, the general tone of reports changed after the 2008 global financial crisis, although subjectivity did not show statistically significant variation. The study highlights that incorporating textual analysis and soft information can enhance the understanding of sovereign credit ratings and underscores the importance of qualitative judgment alongside traditional economic indicators.

The study by Fei, Gu, Yang, and Zhou (2015) examined the effectiveness of social media opinions in predicting the future credit risk of businesses. Traditional financial statement-based logit model results were used as a benchmark. Posts and comments from two social media platforms heavily used by financial investors in China were analyzed using text mining. Opinions of financial analysts were also included in the evaluation. The findings revealed that opinions obtained from social media were more successful than analyst opinions in predicting credit risk and contained meaningful value-related information. This study contributes to the literature by demonstrating that big data and textual analysis techniques can serve as alternatives to traditional methods in credit risk assessment.

In their study, Gül, Kabak, and Topcu (2018) proposed a multi-criteria approach integrating social media data into the credit rating process. In addition to traditional financial and non-financial indicators, data obtained from Twitter was processed using sentiment analysis, criterion weights were determined through pairwise comparisons, and company performance was evaluated using a cumulative belief rating approach. This method provided analysts with a more flexible and interpretable assessment by presenting credit ratings as a risk distribution rather than a single definitive value. The model, applied to 64 companies, revealed that social media data significantly contributed to creditworthiness assessments, but that credit ratings generally tended to decrease when social media data were taken into account.

The study by Chen and Chen (2022) examined the use of social media-derived big data in predicting corporate credit ratings. As an alternative to traditional financial statements, corporate governance, and macroeconomic indicators, they proposed integrating public perception of companies on social media into the credit rating process. In this context, a credit rating prediction process based on social media big data was designed, different machine learning techniques were developed, and the model's application and performance were evaluated. The findings revealed that predictions based on social media data provided higher accuracy compared to traditional indicators, and the K-Nearest Neighbor (KNN) algorithm, in particular, showed superior performance compared to other methods. The study makes a significant contribution to the literature by demonstrating that alternative data sources can be effectively used in credit rating processes within the FinTech context.

In Aksoy's (2020) study, the credit ratings of 11 insurance companies operating in Türkiye's non-life insurance sector with high market share were predicted using machine learning methods. Analyses were performed using Artificial Neural Networks (ANN), k-Nearest Neighbors, and Naive Bayes algorithms with financial statement data from the 2009–2019 period. Prediction performance was evaluated with 10-fold cross-validation. The results showed that ANN achieved a classification success rate of 98.55%, KNN 95.65%, and Naive Bayes 85.51%. Highlighting that these models are effective prediction tools that can be used by insurance companies, lenders, reinsurance companies, and regulatory bodies.

In the study conducted by Doğan, Büyükkör, and Atan (2022), the credit ratings of 1881 companies operating in three different sectors in Türkiye were estimated using machine learning and modern statistical methods. The study observed that logistic regression, support vector machines, Random Forest, and XGBoost algorithms achieved higher classification accuracy, sensitivity, specificity, and precision than decision tree and k-nearest neighbor methods. Furthermore, sector-based analyses were found to significantly affect the performance of credit rating prediction. This study demonstrates that machine learning methods can be used as a reliable and transparent tool in corporate credit ratings in Türkiye.

Sugozu, Verberi, and Yasar (2025) examine the credit risk of participation and traditional banks in Türkiye using machine learning methods. In this study, credit risk models were developed using CatBoost, XGBoost, Random Forest, and LightGBM algorithms with data from 33 traditional and 6 participation banks for the period 2009–2022, and the effects of variables were analyzed using the Tree SHAP method. The findings show that credit risk is higher in participation banks compared to traditional banks, competition increases credit risk, and loan size and profitability play a significant role in risk. Economic growth, however, reduces risk. The study suggests that participation banks should increase their economies of scale and reduce their risks through special regulations. This research makes a significant contribution to the literature by examining the impact of competition on credit risk in Türkiye's dual banking system using interpretable machine learning.

Method

This study adopts a quantitative text analytics design based on sentiment analysis and supervised machine learning. Sentiment analysis is a natural language processing (NLP) and text mining approach that aims to systematically identify subjective expressions in textual data and classify them according to their emotional orientation (positive, negative, neutral; like, dislike) and/or emotional intensity (Saberi & Saad, 2017). One of the pioneering works in the field, Bo Pang and Lillian Lee (2008), define sentiment analysis as the process of automatically extracting and classifying opinions, attitudes, and evaluations. Bing Liu (2012) offers a more comprehensive framework, defining sentiment analysis as the extraction of individuals' opinions and emotions about a particular entity, topic, or event from textual data. In the current literature, this field has evolved into multi-level (sentence, document, orientation/feature-based) analyses that more accurately capture contextual meaning through deep learning, particularly transformer-based models (Jacob Devlin et al., 2019). In this context, thanks to transfer learning, language models previously trained on large-scale datasets can be adapted to more limited, domain-specific datasets. Furthermore, Large Language Models (LLMs) capture contextual meaning, implicit sentiment tones, and complex discourse structures with higher accuracy, thus increasing both the accuracy and generalizability of sentiment analysis (Prottasha et al., 2022).

The dataset consists of corporate credit rating reports that include narratives under *strengths* and *constraints*, along with firm identifiers, sector information, and assigned national rating categories. In this context, the classification of expressions in the textual content as *strengths* and *constraints* is theoretically based on a polarity-based sentiment analysis approach. Within this framework, the "strengths" category represents positive evaluations and elements that strengthen institutional capacity, while the "constraints" category points to negative orientations, limiting factors, and risk indicators. Therefore, the classification performed can be conceptualized as a sentiment analysis application that systematically separates subjective evaluations in texts by sentiment orientation.

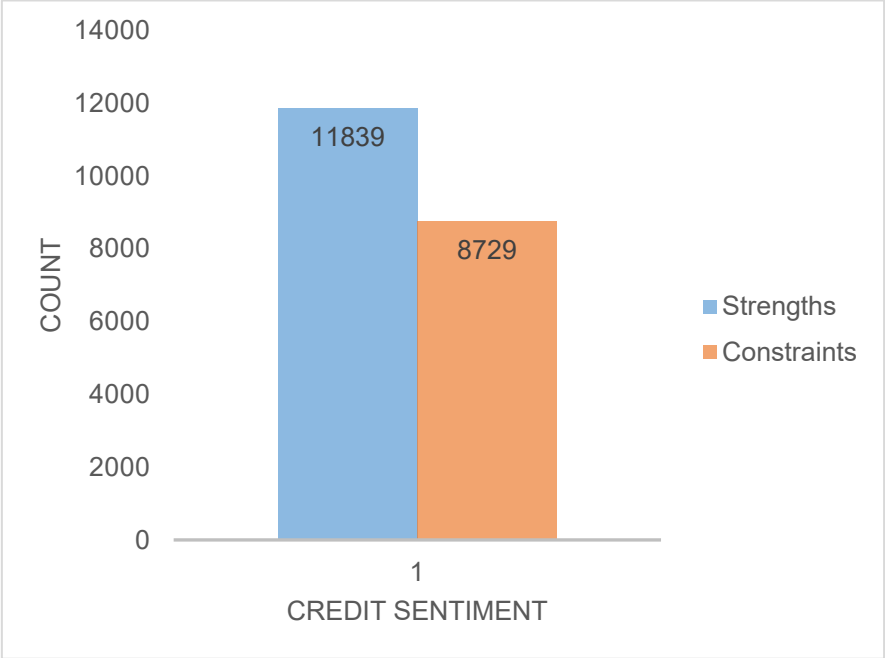
In the first stage, narratives are pre-processed (e.g., normalization, punctuation cleaning, and whitespace standardization) and used to fine-tune a transformer-based classifier to distinguish *strengths* from *constraints*. In the second stage, the model outputs are aggregated to construct report-level indicators, which are then used to predict the national credit rating using a multi-class classification model under a company-level train–test split to prevent information leakage across firms. Because the study relies solely on publicly available rating reports and does not require any data collection from human participants, ethical committee approval was not required.

Data Set

In the study, we used corporate credit rating reports prepared by credit rating agencies as the dataset. These reports generally contain two main types of information. The first consists of textual sections that describe the strengths and weaknesses of firms. In other words, the sections of credit rating reports that express firms' strengths and weaknesses provide a dataset labeled as *strengths* and *constraints* by experts. Within this context, 20,568 labeled observations were obtained from credit rating reports. As shown in Figure 1, 11,839 of these observations correspond to *strengths*, while 8,729 correspond to *constraints*.

Figure 1

Label distribution in the 20k expert-annotated credit sentiment datasets



Note. This figure was created by the author via Python.

The second type of data obtained from the reports consists of the credit ratings assigned to firms as a result of the evaluation process. As is well known, credit ratings are an assessment method based on a firm's ability to meet its financial obligations. They provide a forward-looking projection of the firm's potential to fulfill its financial liabilities in a timely manner.

Credit ratings provide a standardized, transparent language that facilitates global comparability, enabling investors to assess the likelihood that a firm or issuer will repay its debt obligations in full and on time. The rating scale used to evaluate firms and issuances is expressed through the categories "AAA" to "BBB" (Investment Grade), "BB" to "C" (Speculative), and "D" (Default). The use of plus (+) and minus (−) signs allows for further differentiation within the AA to CCC categories.

These categories do not imply that any security is recommended or approved for investment purposes. Ratings in the "Investment Grade" category indicate relatively low to moderate credit risk, whereas those in the "Speculative" category signal higher levels of credit risk. Finally, "Default Event" ratings indicate that a default event has partially or fully occurred.

Evaluation metrics

Model performance was evaluated using standard multi-class classification metrics (Powers, 2011; Sokolova & Lapalme, 2009) reported at both the class and overall levels. Precision measures the proportion of observations predicted as a given class that are correctly classified, while recall measures the proportion of true observations of that class that are correctly identified. Here, TP, FP, and FN denote true positives, false positives, and false negatives, respectively. Accordingly, precision and recall are defined as:

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

The F1-score (Bishop, 2006) is the harmonic mean of precision and recall and provides a single measure that balances these two aspects, particularly useful when there is a trade-off between false positives and false negatives:

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{3}$$

Support denotes the number of test observations belonging to each class (n_k) and is critical for interpreting class-wise results, as metrics computed on very small supports can be unstable. At the aggregate level, accuracy is reported as the fraction of all test observations that are correctly classified:

$$Accuracy = \frac{\sum_{k=1}^{K} TP_k}{N} \tag{4}$$

where N is the total number of test observations and K is the number of classes. In addition, macro-averaged precision/recall/F1 are computed as the unweighted mean across classes, treating each class equally and thus reflecting performance under class imbalance:

$$Macro - M = \frac{1}{K} \sum_{k=1}^{K} M_k \tag{5}$$

By contrast, weighted-averaged metrics weight each class by its support, providing an overall summary that is more influenced by majority classes:

$$\text{Weighted} - \text{M} = \sum_{k=1}^{K} \frac{n_k}{N} M_k \tag{6}$$

Together, these metrics offer a comprehensive view of predictive performance, highlighting both overall correctness and class-specific discrimination.

Implementation Details

Creating sentimental strengths-constraints classifier models

In addition to finance-domain pre-trained models, we also employ other pre-trained large language models with the aim of developing classifier models that detect sentiment in the textual sources of credit rating reports within the *strengths–constraints* framework. Overall, the results indicate strong model performance, with accuracy exceeding 97%. These results remain

consistent across different hyperparameter settings, reinforcing the superiority of pre-trained models.

As shown in Table 1, the performance metrics demonstrate the suitability of the BERT and DistilBERT models for *strengths–constraints*–based sentiment analysis required in credit rating assessment. In the table, the accuracies of the environmental, social, and governance models are reported together with their standard deviations (std.) in parentheses.

Table 1

Strengths–Constraints Classification Results

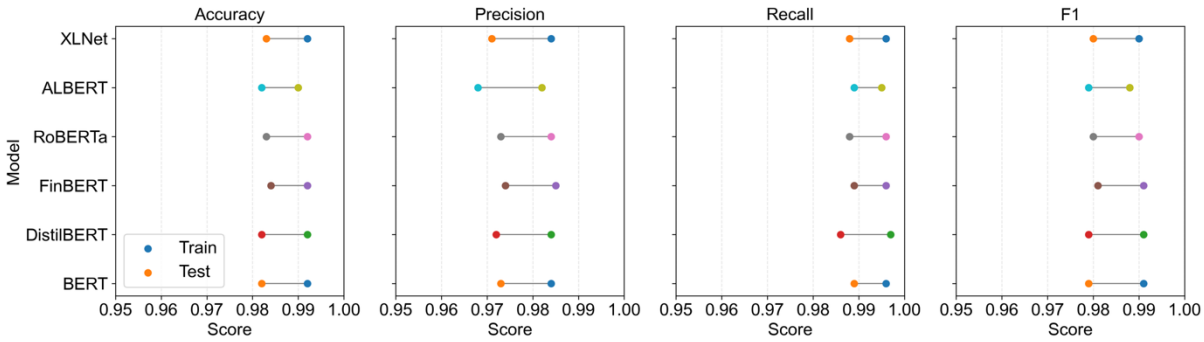
Model	checkpoint	Learning rate	Batch size	epochs	Accuracy		Precision		Recall		F1	
					Train	Test	Train	Test	Train	Test	Train	Test
BERT	bert-base-uncased	0.00002	8	3	0.992	0.982	0.984	0.973	0.996	0.984	0.990	0.978
BERT	bert-base-uncased	0.00003	16	4	0.992	0.982	0.984	0.969	0.997	0.989	0.991	0.979
RoBERTa	roberta-base	0.00002	8	3	0.991	0.984	0.983	0.974	0.996	0.988	0.989	0.981
RoBERTa	roberta-base	0.00003	16	4	0.992	0.983	0.984	0.973	0.996	0.986	0.990	0.980
DistilBERT	distilbert-base-uncased	0.00002	8	3	0.992	0.981	0.985	0.969	0.996	0.986	0.990	0.977
DistilBERT	distilbert-base-uncased	0.00003	16	4	0.992	0.982	0.984	0.972	0.997	0.986	0.991	0.979
ALBERT	albert-base-v2	0.00002	8	3	0.990	0.982	0.982	0.968	0.995	0.989	0.988	0.979
ALBERT	albert-base-v2	0.00003	16	4	0.989	0.973	0.981	0.963	0.993	0.975	0.987	0.969
XLNet	xlnet-base-cased	0.00002	8	3	0.991	0.982	0.983	0.972	0.996	0.986	0.989	0.979
XLNet	xlnet-base-cased	0.00003	16	4	0.992	0.983	0.984	0.971	0.996	0.988	0.990	0.980
FinBERT	ProsusAI/finbert	0.00002	8	3	0.991	0.983	0.983	0.972	0.996	0.989	0.990	0.981
FinBERT	ProsusAI/finbert	0.00003	16	4	0.992	0.984	0.985	0.974	0.996	0.988	0.991	0.981

Note. Created by the authors using Python.

Table 1 reports the performance of different Transformer-based language models on the *strengths–constraints* binary classification task. For BERT, RoBERTa, DistilBERT, ALBERT, XLNet, and FinBERT, test accuracy rates exceed 97%, while test F1 scores are above 0.97.

Figure 2

Train vs. Test Performance Across Models and Metrics

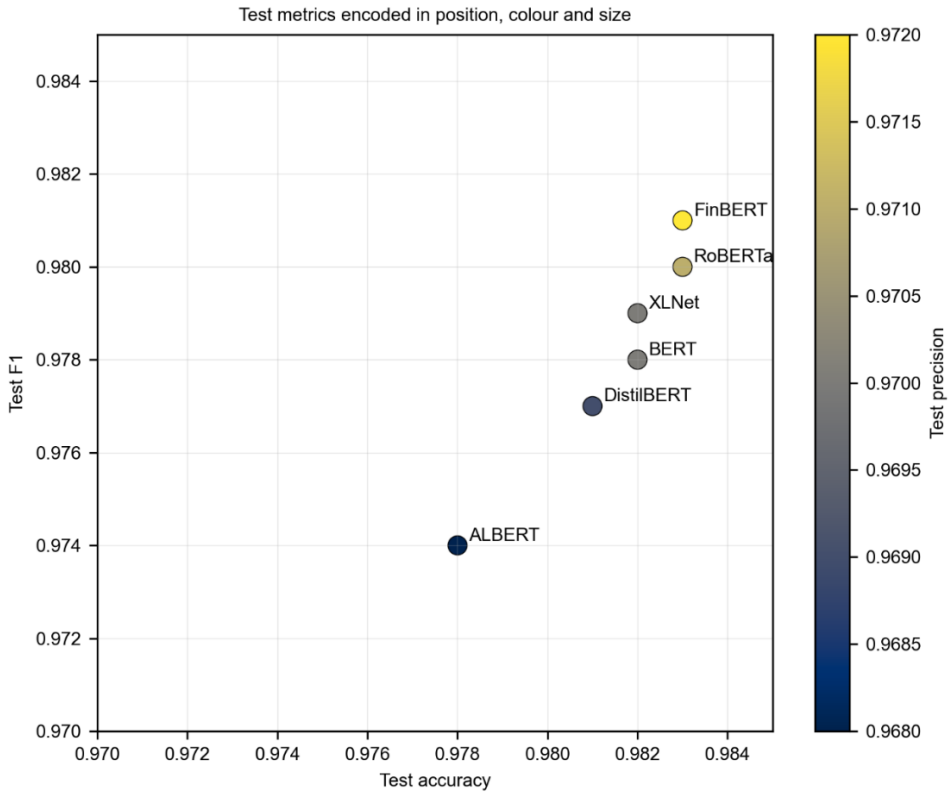


Note. This figure was created by the author via Python, where Arial font customization is not supported.

As shown in Figure 2, the fact that the gap between training and test metrics remains below 0.01 across all models indicates strong generalization capability. The highest test F1 score, 0.981, is achieved by the RoBERTa-base and FinBERT configurations. The results suggest that FinBERT's pre-training on financial text provides a limited but observable performance advantage in the context of credit rating reports.

Figure 3

Performance Results of Language Models

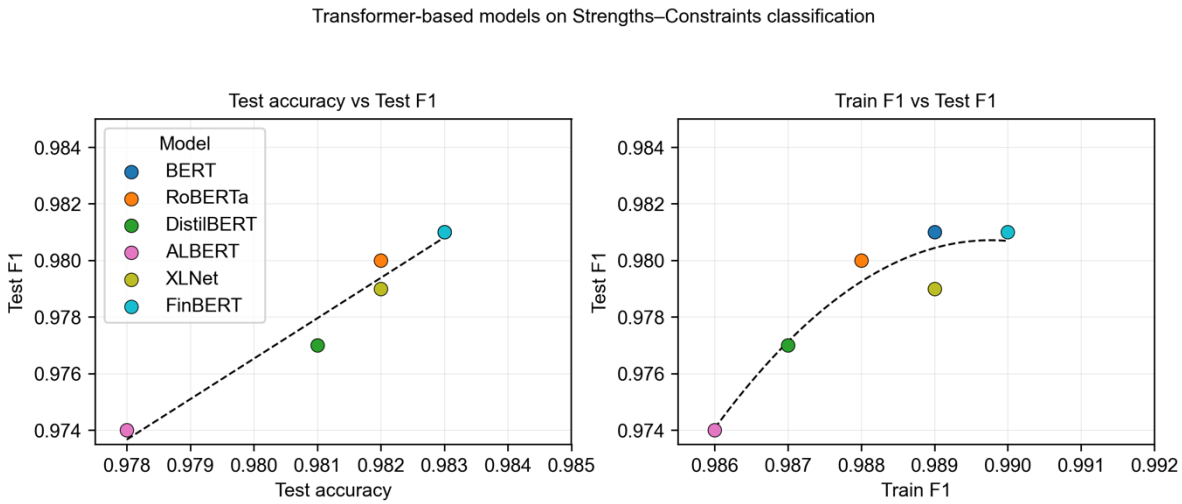


Note. This figure was created by the author via Python, where Arial font customization is not supported.

Nevertheless, the relatively small difference between general-purpose BERT-derived models and the domain-specific FinBERT indicates that the *strengths* and *constraints* sections are linguistically well separated and that the task is similarly tractable across different Transformer architectures. Moreover, the ability of the lighter DistilBERT architecture to produce results very close to those of larger models can be considered an important finding in terms of the trade-off between computational cost and performance at the deployment stage (Figure 3).

Figure 4

Transormer-Based Models on Strengths-Constraints Classification



Note. This figure was created by the author via Python, where Arial font customization is not supported.

As shown in Figure 4, test accuracy and test F1 scores (left) and train–test F1 scores (right) for different hyperparameter configurations of Transformer-based models in the *strengths–constraints* binary classification task. Each point represents a specific model–hyperparameter combination. The orange lines indicate smoothed curves obtained using a second-degree polynomial. The axes are zoomed into the 0.97–0.985 range. Therefore, differences across models are very small in absolute terms and are shown solely for relative comparison purposes.

From strengths-constraints to credit rating classifier models

Two different approaches are adopted to develop a model that predicts firms' credit ratings based on reports issued by credit rating agencies. The first approach classifies texts into *strengths* and *constraints*, then derives a sentiment score from this classification and uses that score to predict firms' credit ratings.

In the second approach, *strengths–constraints* texts are converted into vectors using Transformer libraries, thereby reducing all information to a single vector representation and embedding each text in a semantic vector space. In other words, if *strengths* are considered as positive aspects and *constraints* as negative aspects, these two types of text complement each other and together form a unified representation of a firm's overall financial condition. By providing *strengths* and *constraints* texts jointly to the model, it is ensured that the model learns the context of both sides within the same embedding space.

Table 2

Rating Scale

Investment Level	Rating Group	Credit Quality	Remarks
Investment Grade	AAA	Highest Credit Quality	Represents the minimal expected risk of default. Assigned only in exceptional circumstances where the entity demonstrates a very strong ability to meet its financial obligations, with this capacity being highly resilient to foreseeable events.
	AA	Very High Credit Quality	Reflects a very low expected level of default risk. It is assigned in exceptional situations where the entity demonstrates a strong ability to meet its financial obligations, and this ability is highly resilient to foreseeable events.
	A	High Credit Quality	Indicates a low expected level of default risk. The entity is considered to have a strong ability to meet its financial obligations. However, this ability may be more susceptible to negative business or economic developments compared with higher-rated entities.
	BBB	Good Credit Quality	Indicates that the expected default risk is presently low. The entity's capacity to meet financial obligations is considered adequateh. However, unfavorable business or economic conditions are more likely to compromise its solvency.
Speculative	BB	Speculative	Indicates a higher susceptibility to default risk, especially if business or economic conditions deteriorate over time. Nevertheless, the entity retains commercial or financial flexibility that helps ensure fulfillment of its financial obligations.
	B	Highly Speculative	Indicates a significant default risk, with only a narrow margin of safety. Although financial obligations are currently being met, the entity's ability to fulfill them remains vulnerable to adverse changes in business and economic conditions.

Note. Adapted by the author from "Credit Rating" (https://www.jcrer.com.tr/en/methodology/notations/credit-rating).

Table 2 presents a structured classification of credit ratings, linking investment levels, rating groups, and corresponding credit quality to the expected default risk and the entity's capacity to meet financial obligations. Investment-grade ratings, ranging from AAA to BBB, reflect minimal to low default risk, indicating a very strong to adequate ability to fulfill financial commitments, with resilience varying by rating. AAA and AA ratings represent the highest and very high credit quality, assigned in exceptional circumstances in which the entity demonstrates a strong, highly resilient capacity to meet its obligations. A and BBB ratings denote high to good credit quality, with the ability to meet obligations considered strong to adequate, though more susceptible to adverse business or economic developments compared to higher-rated entities. Speculative ratings, including BB and B, indicate elevated to significant default risk, where financial obligations may still be met but are increasingly vulnerable to deteriorating business and economic conditions (Category B was combined with Category BB in the analysis because there were very few observations in category B.). The table highlights not only the expected probability of default but also the degree of financial flexibility and resilience associated with each rating category, providing a comprehensive framework for evaluating credit quality across different investment levels.

Table 3

Credit Score Classification Results

Credit Class	Precision	Recall	F1-score	Support
AAA	0.4000	0.1081	0.1702	37
AA	0.5669	0.6486	0.6050	111
A	0.4202	0.4808	0.4484	104
BBB	0.5851	0.5978	0.5914	92
BB	0.5000	0.1250	0.2000	8
accuracy			**0.5170**	352
macro avg	0.4944	0.3921	0.4030	352
weighted avg	0.5093	0.5170	0.5003	352

Note. Created by the authors using Python.

The results indicate that narrative-based classification and credit rating prediction involve different levels of difficulty. In the *strength–constraint* stage (Task 1), the transformer model converged with low training loss and achieved strong test performance. In contrast, in Task 2, where the national credit rating was predicted using only report-level narrative features derived from Task1, performance was more limited (accuracy = 0.517; macro-F1 = 0.403; weighted-F1 = 0.500; n = 352). A class-wise evaluation shows relatively more stable performance for the AA (F1 = 0.605; recall = 0.649) and BBB (F1 = 0.591; recall = 0.598) categories, while the model attains moderate success for A (F1 = 0.448). By comparison, the markedly low recall for AAA (recall = 0.108; F1 = 0.170) suggests substantial difficulty in capturing the highest rating category. Similarly, due to the very small support for BB (support = 8), the corresponding metrics are unstable (recall = 0.125; F1 = 0.200). Overall, these findings suggest that indicators derived from report narratives contain meaningful signals about assigned ratings. However, when used alone, predictive power remains constrained, particularly for extreme classes and under pronounced class imbalance.

Discussion

This study examines the predictive performance of transformer-based language models for credit score classification. All evaluated models (BERT, RoBERTa, DistilBERT, ALBERT, XLNet, and FinBERT) showed very high performance on both training and test accuracy in the *strengths–constraints* classification task, achieving F1 scores above 0.97. This demonstrates that transformer architectures can effectively capture the necessary textual and financial information to identify *strengths* and *constraints* in corporate credit assessments, confirming the potential of deep learning approaches in financial risk assessment.

These findings support the work of Chen et al. (2022) and Fei et al. (2015), which showed that big data from social media and text analysis techniques can improve accuracy in credit risk prediction. In this study, transformer-based models were applied to structured and semi-structured textual data rather than social media data, and similarly high predictive performance was achieved. This situation parallels the general findings in the literature that

alternative textual data can improve the accuracy of credit rating processes (Gül, Kabak, & Topcu, 2018; Slapnik & Lončarski, 2023).

Unlike *strengths-constraints* classification, credit score classification proved to be a more challenging task. The weighted average accuracy was 0.517, and significant differences were observed across credit classes (e.g., AAA F1 score: 0.17, BBB F1 score: 0.59). This reflects the difficulty of distinguishing between high-grade and low-frequency classes. In this context, the low number of observations in some classes (e.g., 8 samples for the BB class) may have negatively impacted model performance.

Methodologically, FinBERT consistently demonstrated high performance in *strengths-constraints* tasks. This is consistent with the FinTech literature supporting the value of using domain-specific pre-trained models in financial NLP applications (Sugozu et al., 2025). The small performance differences between BERT, RoBERTa, and DistilBERT demonstrate that basic transformer models can generalize well, but domain adaptation and fine-tuning are critical to maximizing prediction accuracy. Overall, these results offer several important insights. First, transformer-based models can accurately predict performance-related constraints by extracting strong representations from financial texts. Second, credit rating predictions are sensitive to overclasses and limited sample sizes. Third, integration with text analyses from social media and credit reports (Slapnik & Lončarski, 2023; Fei et al., 2015) offers a promising avenue for improving predictive models.

Conclusion

This study makes a significant contribution to credit analytics by demonstrating that qualitative narratives in credit rating reports, which constitute the most interpretation-intensive component of the rating process, can be transformed into consistent quantitative information and integrated into an end-to-end modeling pipeline. First, the near-ceiling performance of the *strength–constraint* classifier provides strong evidence that rating reports contain stable and learnable linguistic patterns. This, in turn, enables the scalable and automated extraction of positive and negative rating drivers. More broadly, the findings point to a practical mechanism that standardizes the interpretation of narratives, reduces reliance on manual reading, and produces auditable summaries of credit rationales across firms and sectors.

Second, the rating prediction analysis delivers a clear empirical insight into the informational value of narratives. Even when the model is restricted to report-level features extracted from text, it can meaningfully distinguish among the more frequently observed rating categories, suggesting that qualitative assessments contain a non-trivial signal related to the final rating decision. At the same time, the weaker results for rare and extreme categories, most notably the difficulty in identifying AAA and the unreliability of BB estimates given its very small sample size, offer an important diagnostic implication. These patterns indicate where narrative information alone is insufficient and where factors such as class imbalance, threshold dynamics, and the inclusion of complementary non-textual inputs become crucial. Accordingly, the study clarifies the practical limits of text-only credit modeling rather than implying that narratives can fully substitute for the broader information set used in rating assignments.

From an applied standpoint, the proposed framework provides a scalable foundation for continuous credit monitoring. It can be deployed to automatically parse newly released rating reports, quantify shifts in the balance of *strengths* and *constraints*, and generate early-warning indicators for analysts, investors, and risk managers. Conceptually, the study bridges qualitative credit reasoning and quantitative prediction by operationalizing narrative content into transparent indicators that can be validated, compared over time, and integrated with other

information sources. Future research can build on this foundation by combining narratives with richer quantitative fundamentals and macro signals, adopting ordinal/hierarchical and cost-sensitive learning to better reflect the rating scale, and testing robustness across time periods and rating agencies to strengthen real-world generalizability.

References

Aksoy, B. (2020). Sigorta şirketlerinin derecelendirilmesinde makine öğrenmesi yöntemleri tahmin performansının karşılaştırılması: Türkiye örneği. *Akademik Araştırmalar ve Çalışmalar Dergisi (AKAD)*, 12(23), 579-597. https://doi.org/10.20990/kilisiibfakademik.710863

Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J., & Vanthienen, J. (2003). Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the Operational Research Society*, *54*(6), 627–635. https://doi.org/10.1057/palgrave.jors.2601545

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Boughaci, D., & Alkhawaldeh, A. A. K. (2020). Appropriate machine learning techniques for credit scoring and bankruptcy prediction in banking and finance: A comparative study. *Risk and Decision Analysis*, *8*(1–2), 15–24. https://doi.org/10.3233/RDA-180051

Bucker, M., Szepannek, G., Gosiewska, A., & Biecek, P. (2022). Transparency, auditability, and explainability of machine learning models in credit scoring. *Journal of the Operational Research Society, 73*(1), 70–90. https://doi.org/10.1080/01605682.2021.1922098

Chen, Y.-J., & Chen, Y.-M. (2022). Forecasting corporate credit ratings using big data from social media. *Expert Systems with Applications, 207*, Article 118042. https://doi.org/10.1016/j.eswa.2022.118042

Darwish, J. A. (2025). Optimization and prediction of corporate credit rating through advanced feature selection based on AI and deep learning. *Alexandria Engineering Journal*, *127*, 586–594. https://doi.org/10.1016/j.aej.2025.05.043

De Oliveira, N. A., & Basso, L. F. C. (2025). Advancing credit rating prediction: The role of machine learning in corporate credit rating assessment. *Risks*, *13*(6), 116. https://doi.org/10.3390/risks13060116

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long and Short Papers)* (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Doğan, S., Büyükkör, Y., & Atan, M. (2022). A comparative study of corporate credit ratings prediction with machine learning. *Operations Research and Decisions, 32*(1), 25–47. https://doi.org/10.37190/ord220102

Dumitrescu, E., Hué, S., Hurlin, C., & Tokpavi, S. (2022). Machine learning for credit scoring: Improving logistic regression with non-linear decision-tree effects. *European Journal of Operational Research*, *297*(3), 1178–1192. https://doi.org/10.1016/j.ejor.2021.06.053

Fei, W., Gu, J., Yang, Y., & Zhou, Z. (2015). Credit risk evaluation based on social media. *Procedia Computer Science, 55*, 725–731. https://doi.org/10.1016/j.procs.2015.07.165

Gajdosikova, D., & Valaskova, K. (2023). Bankruptcy prediction model development and its implications on financial performance in Slovakia. *Economics and Culture*, *20*(1), 30–42. https://doi.org/10.2478/jec-2023-0003

Gambacorta, L., Huang, Y., Qiu, H., & Wang, J. (2024). How do machine learning and non-traditional data affect credit scoring? New evidence from a Chinese fintech firm. *Journal of Financial Stability*, *73*, Article 101284. https://doi.org/10.1016/j.jfs.2024.101284

Gunnarsson, B. R., vanden Broucke, S., Baesens, B., Óskarsdóttir, M., & Lemahieu, W. (2021). Deep learning for credit scoring: Do or don't? *European Journal of Operational Research*, *295*(1), 292–305. https://doi.org/10.1016/j.ejor.2021.03.006

Gül, S., Kabak, Ö., & Topcu, I. (2018). A multiple criteria credit rating approach utilizing social media data. *Data & Knowledge Engineering, 116*, 80–99. https://doi.org/10.1016/j.datak.2018.05.005

Issa, S., Bizel, G., Jagannathan, S. K., & Gollapalli, S. S. C. (2024). A comprehensive approach to bankruptcy risk evaluation in the financial industry. *Journal of Risk and Financial Management*, *17*(1), 41. https://doi.org/10.3390/jrfm17010041

Lessmann, S., Baesens, B., Seow, H.-V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, *247*(1), 124–136. https://doi.org/10.1016/j.ejor.2015.05.030

Liu, B. (2012). *Sentiment analysis and opinion mining*. Morgan & Claypool Publishers.

Liu, Y., & Schumann, M. (2005). Data mining feature selection for credit scoring models. *The Journal of the Operational Research Society*, *56*(9), 1099–1108. https://doi.org/10.1057/palgrave.jors.2601976

Machado, M. R., & Karray, S. (2022). Assessing credit risk of commercial customers using hybrid machine learning algorithms. *Expert Systems with Applications*, *200*, Article 116889. https://doi.org/10.1016/j.eswa.2022.116889

Machado, M. R., Chen, D. T., & Osterrieder, J. R. (2025). An analytical approach to credit risk assessment using machine learning models. *Decision Analytics Journal*, *16*, Article 100605. https://doi.org/10.1016/j.dajour.2025.100605

Marqués, A. I., García, V., & Sánchez, J. S. (2012). Two-level classifier ensembles for credit risk assessment. *Expert Systems with Applications*, *39*(12), 10916–10922. https://doi.org/10.1016/j.eswa.2012.03.033

Máté, D., Raza, H., & Ahmad, I. (2023). Comparative analysis of machine learning models for bankruptcy prediction in the context of Pakistani companies. *Risks*, *11*(10), Article 176. https://doi.org/10.3390/risks11100176

Muñoz-Izquierdo, N., Segovia-Vargas, M. J., Camacho-Miñano, M.-M., & Pérez-Pérez, Y. (2022). Machine learning in corporate credit rating assessment using the expanded audit report. *Machine Learning*, *111*(11), 4183–4215. https://doi.org/10.1007/s10994-022-06226-4

Pai, P.-F., Tan, Y.-S., & Hsu, M.-F. (2015). Credit rating analysis by the decision-tree support vector machine with ensemble strategies. *International Journal of Fuzzy Systems*, *17*(4), 521–530. https://doi.org/10.1007/s40815-015-0063-y

Paleologo, G., Elisseeff, A., & Antonini, G. (2010). Subagging for credit scoring models. *European Journal of Operational Research*, *201*(2), 490–499. https://doi.org/10.1016/j.ejor.2009.03.008

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, *2*(1–2), 1–135. https://doi.org/10.1561/1500000001

Powers, D. M. W. (2011). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *Journal of Machine Learning Technologies, 2*(1), 37–63. https://doi.org/10.9735/2229-3981

Prottasha, N. J., Sami, A. A., Kowsher, M., Murad, S. A., Bairagi, A. K., Masud, M., & Baz, M. (2022). Transfer learning for sentiment analysis using BERT based supervised fine-tuning. *Sensors*, *22*(11), 4157. https://doi.org/10.3390/s22114157

Radovanovic, J., & Haas, C. (2023). The evaluation of bankruptcy prediction models based on socio-economic costs. *Expert Systems with Applications*, *227*, Article 120275. https://doi.org/10.1016/j.eswa.2023.120275

Saberi, B., & Saad, S. (2017). Sentiment analysis or opinion mining: A review. .*International Journal of Advanced Science Engineering Information Technology*, *7*(5), 1660-1666.

Siham, A., Sekkate, S., & Adib, A. (2021). Feature selection based on machine learning for credit scoring: An evaluation of filter and embedded methods. In Z. H. Kilimci, T. Yildirim, V. Piuri, I. Czarnowski, D. Camacho, Y. Manolopoulos, & S. Solak (Eds.), *Proceedings of the 2021 International Conference on INnovations in Intelligent SysTems and Applications (INISTA 2021)* (pp. 1–6). IEEE. https://doi.org/10.1109/INISTA52262.2021.9548410

Slapnik, U., & Lončarski, I. (2023). Understanding sovereign credit ratings: Text-based evidence from the credit rating reports. *Journal of International Financial Markets, Institutions and Money, 88*, Article 101838. https://doi.org/10.1016/j.intfin.2023.101838

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, *45*(4), 427–437. https://doi.org/10.1016/j.ipm.2009.03.002

Sugozu, I. H., Verberi, C., & Yasar, S. (2025). Machine learning approaches to credit risk: Evaluating Turkish participation and conventional banks. *Borsa Istanbul Review*, *25*(3), 497-512. https://doi.org/10.1016/j.bir.2025.02.001

Teles, G., Rodrigues, J. J. P. C., Rabêlo, R. A. L., & Kozlov, S. A. (2021). Comparative study of support vector machines and random forests machine learning algorithms on credit operation. *Software: Practice and Experience*, *51*(12), 2492–2500. https://doi.org/10.1002/spe.2842

Teles, G., Rodrigues, J. J. P. C., Saleem, K., Kozlov, S., & Rabêlo, R. A. L. (2020). Machine learning and decision support system on credit scoring. *Neural Computing and Applications*, *32*(14), 9809–9826. https://doi.org/10.1007/s00521-019-04537-7

Thomas, L. C. (2000). A survey of credit and behavioural scoring: Forecasting financial risk of lending to consumers. *International Journal of Forecasting*, *16*(2), 149–172. https://doi.org/10.1016/S0169-2070(00)00034-0

Toudas, K., Archontakis, S., & Boufounou, P. (2024). Corporate bankruptcy prediction models: A comparative study for the construction sector in Greece. *Computation*, *12*(1), 9. https://doi.org/10.3390/computation12010009

Tripathi, D., Edla, D. R., Cheruku, R., & Kuppili, V. (2019). A novel hybrid credit scoring model based on ensemble feature selection and multilayer ensemble classification. *Computational Intelligence*, *35*(2), 371–394. https://doi.org/10.1111/coin.12200

Tripathi, D., Edla, D. R., Kuppili, V., & Bablani, A. (2020). Evolutionary extreme learning machine with novel activation function for credit scoring. *Engineering Applications of Artificial Intelligence*, *96*, Article 103980. https://doi.org/10.1016/j.engappai.2020.103980

Trivedi, S. K. (2020). A study on credit scoring modeling with different feature selection and machine learning approaches. *Technology in Society*, *63*, Article 101413. https://doi.org/10.1016/j.techsoc.2020.101413

Tsai, C.-F. (2014). Combining cluster analysis with classifier ensembles to predict financial distress. *Information Fusion*, *16*, 46–58. https://doi.org/10.1016/j.inffus.2011.12.001

Wang, G., Hao, J., Ma, J., & Jiang, H. (2011). A comparative assessment of ensemble learning for credit scoring. *Expert Systems with Applications*, *38*(1), 223–230. https://doi.org/10.1016/j.eswa.2010.06.048

Wu, Y., & Pan, Y. (2021). Application analysis of credit scoring of financial institutions based on machine learning model. *Complexity*, *2021*(1), Article 9222617. https://doi.org/10.1155/2021/9222617

Xu, Y., Chen, Y., Sun, L., & Chen, Y. (2025). Corporate credit scoring method based on unlabeled data and multi-source data. *Decision Support Systems*, *198*, Article 114543. https://doi.org/10.1016/j.dss.2025.114543

Yobas, M. B., Crook, J. N., & Ross, P. (2000). Credit scoring using neural and evolutionary techniques. *IMA Journal of Management Mathematics*, *11*(2), 111-125. https://doi.org/10.1093/imaman/11.2.11

Information About the Article/Makale Hakkında Bilgiler

The Ethical Rules for Research and Publication / Araştırma ve Yayın Etiği

The author declared that the ethical rules for research and publication followed while preparing the article.

Yazar makale hazırlanırken araştırma ve yayın etiğine uyulduğunu beyan etmiştir.

Conflict of Interests/ Çıkar Çatışması

The author have no conflict of interest to declare.

Yazar çıkar çatışması bildirmemiştir.

Grant Support/ Finansal Destek

The author declared that this study has received no financial support.

Yazar bu çalışma için finansal destek almadığını beyan etmiştir.

Author Contributions/ Yazar Katkıları

The draft process of the manuscript/ Taslağın Hazırlanma Süreci Y.E.A, Data Collection/Verilerin Toplanması Y.E.A, Writing the Manuscript/ Makalenin Yazılması Y.E.A, Submit, Revision and Resubmit Process/ Başvuru, Düzeltme ve Yeniden Başvuru Süreci Y.E.A.